Genetic Adaptive Neural Networks for Prediction of Insurance Claims

Fred Kitchens, Thomas Harris

***Abstract*— In the insurance business, an underwriter's two most important considerations are loss frequency and loss severity (probability of a loss and the financial value of the loss). Neural networks have been successfully applied to the insurance business in areas such as prediction of loss frequency, and prediction of bankruptcy. The objective of this study is to develop a neural network model to predict the severity of potential insurance losses in private passenger automobile insurance in the United States. The study predicts loss severity in private passenger automobile insurance using independent variables commonly available to the insurance underwriter based on the consumer's application for insurance. A genetic adaptive neural network training algorithm is used to model the losses. Premiums are expected to be positively correlated with risk; therefore a linear model is developed as a benchmark, using the same data. These findings show that loss severity is more difficult to predict than frequency of loss. Improved results are likely to be found through alternative methodology, or data.**

***Index Terms*— Automobile Insurance, GANNT, Genetic Adaptive Neural Network Training Algorithm, Linear Regression.**

I. Introduction

Today's insurance market is predicated on the notion that an insurance underwriter is able to scrutinize a set of applications for insurance and select a subset for which the collective premium will be greater than the collective losses. This free-market insurance system dates back to the early 1300's and the founding of Lloyd's of London [1],[2].

Traditionally, the underwriter evaluated each case based on what he considered loss-contributing factors. The underwriter's experience in the insurance business, life, and education are all considered important in his ability to evaluate an insurance policy. The greater his cumulative experience, the greater his anticipated understanding of the relationships between the policy characteristics and the potential loss frequency and severity. In theory, if an experienced underwriter performs his job well, the premium on each policy will have a linear relationship to the risk associated with the policy. Further, in the insurance business, risk is associated with financial losses; therefore, the premium and losses are expected to have a positive linear correlation.

As the field of insurance developed over time, insurance companies hired actuaries to assist the underwriter in the policy selection and pricing process. The actuary's function is to analyze past policy characteristics and loss experiences to find relationships that the underwriter may use in the policy selection and pricing [3]. The actuaries issue underwriting guidelines that, among other purposes; serve to ensure uniform and consistent underwriting among all underwriters for a particular company, and synthesize insights and experiences of experienced underwriters [4].

For many years, the actuaries performed their duties using pencil and paper. Society and the insurance industry evolved to a stage where more advanced computing tools became available [5],[6],[7]. In fact, insurers fell behind other industries in the application of technology to their business processes [8]. One such technology-based tool with potential as an aid to the underwriting process is the artificial neural network [9].

The utility of artificial neural networks has been demonstrated in other fields such as predicting bankruptcy [10],[11], predicting insurance company insolvency [12], analyzing commercial bank failures [13] and predicting farm debt failures [14].

II. Objective

The objective of this study is to develop a neural network model to predict the severity of potential insurance losses in private passenger automobile insurance in the United States. Specifically, the Genetic Adaptive Neural Network Training algorithm (GANNT) will be applied to independent variables commonly available to the insurance underwriter at the time the application for insurance is processed. Premium is expected to be a direct reflection of the risk associated with a given policy. Therefore, a linear model is developed as a benchmark, using the same data.

III. Methods

This study will apply artificial neural networks and linear regression to private passenger automobile insurance policies in a three-step procedure. The results from the two models will be compared using the Wilcoxon Signed Ranks Test.

A. Artificial neural network

Artificial neural networks comprise a class of nonlinear statistical models which process information through a process analogous to the functioning of the human brain [2]. The advantage of neural network models over other methods grows with the complexity of the relationship between inputs and output. The more complex the underlying relationship between variables, the more complex the neural network will need to be (i.e.: more hidden nodes) [15]. Properly designed, a neural network is capable of approximating any unknown function to any degree of accuracy [16].

This study used the Genetic Adaptive Neural Network Training algorithm (GANNT). The GANNT is able to overcome certain problems associated with the popular back-propagation and gradient search techniques [17]. The GANNT solves computer-based problems by modeling a biological process. DNA reproduces itself through a process of separation, crossover, and mutation [18]. In its search for an optimal solution, the GANNT algorithm uses an analogous process as it manipulates arrays of data in search of an optimal

solution [19]. A detailed explanation of the process can be found in recent literature [20].

B. Data

The dataset consists of 174,000 insurance policies from a large insurer of private passenger automobiles in the United States. Records included the insurance application, the drivers' motor vehicle records, the history of past losses for the policy, a record of the losses incurred within the effective dates of each policy, and the premium charged for the insurance coverage. Losses were tracked either to settlement or for two years. After two years, the settlement value was considered equal to the reserved value of the claim.

Automobile insurance in particular was most suitable for this type of study due to the large volume of data available, the relative consistency of the exposure (as compared to other lines of insurance), and the inherent categorical and scalar nature of the data involved.

The variables included one dependent and sixteen independent variables. The dependent variable was the total value of all losses on a policy during the time policy was in force. The independent variables were: Earned premium per exposure unit, Number of at-fault accidents, Number of not-at-fault accidents, Number of convictions, any Restricted vehicles, Vehicle model year - maximum, Vehicle model year - minimum, Number of vehicles, Mileage maximum, Mileage minimum, Age of the youngest driver, Age of the oldest driver, Number of operators (primary only), Number of excess operators, Number of male operators, Number of married operators.

C. Procedures

This study compares two models, a GANNT model and a linear regression model. In theory, if the premium accurately reflects the risk on a given policy, a linear model would fit the data with minimal residual variation.

The linear model uses regression to analyze the extent to which the premium reflects the risk on each policy. The basis of the model is the traditional linear equation:

$$y = a_1 x_1 + a_2 x_2 + ...a_n x_n + b \tag{1}$$

The neural network model uses the sigmoid function. This provides the basis of its flexible and non-linear form:

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{2}$$

The design of this study requires three steps. The first step is to draw random samples of data. Sixty sample sets of data were drawn at random (with replacement); 30 to be used as in-sample data and 30 to be used as out-of-sample data. The accepted rule of thumb in determining sample size is to use at least five cases for every hidden node [22]. If up to 16 hidden nodes might be used (because there are 16 independent variables), the rule of thumb would require at least 80 observations per sample. In this situation, with 174,000 observations available, we were not limited by the availability of data. However, the time requirements to train the neural networks did limit the number of observations chosen to use per sample. For this reason, balanced samples of 500 observations were chosen (250 with losses and 250 without losses).

The second step was to determine the optimal number of hidden nodes to use in the neural network. A neural network was repeatedly trained on the same set of in-sample data, using every possible number of hidden nodes, from 1 to 16.

The third step was to prepare all 30 neural network models and 30 linear regression models using the same sets of in-sample and out-of-sample data, then comparing the results. Results consist of the Root Mean Squared Error (RMSE) for each set of data. Tests of significance can be made using the Wilcoxon Signed Ranks Test, a nonparametric test of the median error values.

IV. Results

After drawing random samples of data, the optimal number of hidden nodes was determined. Sixteen neural networks were trained using the first in-sample data set and tested using the first out-of-sample data set. In testing for the optimal number of hidden nodes, if everything works as anticipated, two things should happen. One, the RMSE of the in-sample data will continue to decrease as the number of hidden-nodes increases. Two, the RMSE of the out-of-sample data will initially decrease as the number of hidden nodes increases. This is due to the improved model configurations rooting-out additional interaction effects between variables. Eventually, at some number of hidden nodes the RMSE of the out-of-sample data will begin to increase. This is an indication that the model has begun to over-fit the in-sample data at the expense of the out-of-sample data's RMSE. Fig. 1 depicts the RMSE values for each of the 16 neural network models.

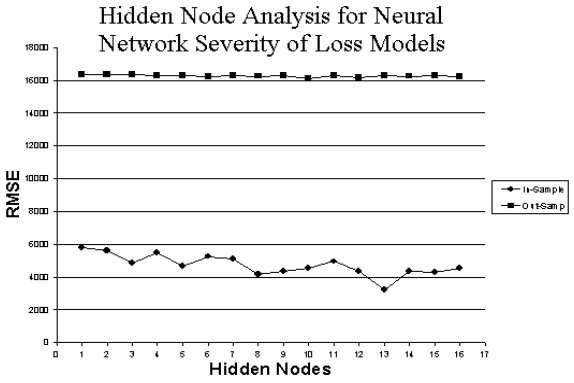


Figure 1: Hidden Node Analysis for Neural Network Severity of Loss Model

In this study, the out-of-sample RMSE values show very little distinction among one another. Although the model with 10 hidden nodes had the lowest out-of-sample RMSE at 16,132 (the in-sample RMSE was 4,542.37), the reduction in value was minimal (the range over all 16 models was a mere 210, with an average of 16,266.19). Given the slight margin of improvement in the 10-hidden node model, we considered it prudent to calculate the RMSE for the linear model using this same data, before proceeding to train the remaining 29 models.

Using the same data, the RMSE values for the in-sample and out-of sample data using a linear regression model were:

In-sample: *5,709.18*

Out-of-sample: *16,410.29*

In both the linear and neural network cases, the difference

between the in- and out-of-sample RMSEs was so great that we considered it an indication that other independent variables should be considered in future models. The differences in the neural network model's RMSE and the linear regression model's RMSE was so slight, we tested the significance of the difference before proceeding with the study.

The Wilcoxon Signed Ranks test was used to test for a significant difference between the results of the neural network model and the linear regression model, Tables 1 and 2 show the calculations and results of this test.

Table 1: Wilcoxon Signed Ranks Test Results

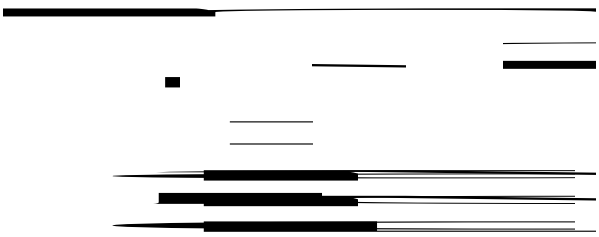
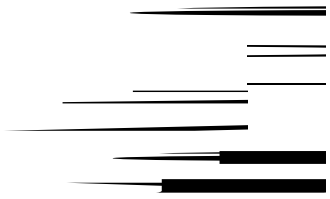


Table 2: Wilcoxon Signed Ranks Test Statistics



These results show that there is a significant difference between the medians of the error terms between the two models. However, the practical significance should also be considered in this situation. Practical significance is a judgment that must be made by the researcher [23]. In this study, the results were found to be statistically significant. However, the significance is based on large sample sizes, meaning that they may not have practical significance. Sometimes, practical significance can be judged considering the means and the range of possible values [24].

In this study, the improvement from the linear model to the neural network model was only 1.69 percent (16,410.29 to 16,132.52). However, the actual losses ranged from $0 (for a loss that was either claimed but payment was denied, or for which the deductible value was not exceeded) to $375,621.56. The average paid claim was $3,823, with a standard deviation of $13,323.60. We determined that while these results showed statistical significance, they were not practically significant.

V. Conclusions

Previous researched found promising results in a similar study designed to predict frequency of loss rather than severity of loss. The difference is that in predicting loss frequency, the dependent variable is dichotomous (0 or 1). In predicting severity, the dependent variable is scalar – in this case ranging from $0.00 to $375,621.56. As it turned out, we underestimated the added difficulty of predicting a scalar dollar value rather than a categorical loss /no-loss.

Further study is expected to provide improved results. However, a reevaluation of the independent variables is warranted. While the quantity of data used in this study was more than sufficient, the type of information contained in the independent variables should be revised.

The incidence of a loss has a great deal to do with both the car and the driver. Characteristics of the driver such as age and previous accidents lend to the maturity and experience of the driver. Characteristics of the vehicle, such as age and mileage may lead to variation in the incident of loss through factors such as poor brakes, and worn tires.

Recent research has indicated that additional data such as GPS could be used to predict short-term travel conditions [25]. The Progressive Insurance Company has implemented the 'Motor Vehicle Monitoring System' for determining a cost of insurance [26],[21],[27].

The severity of a loss is likely to be influenced less by the driver and more by the characteristics of the car. Characteristics such as the make, model, and value of the vehicle were not considered in this study. These characteristics will affect the cost of labor and replacement parts, in addition to the owner's propensity to maintain the car in good working order.

Other factors affecting the value of a claim, and having little to do with the car or driver include medical payments and lawsuits. Medical payments can fluctuate widely and may be based on things unrelated to the car or driver – such as the attending doctors' competence, the administration first-aid, and weather conditions at the time of the accident. The outcome of lawsuits may involve such factors as the lawyers' experience and expertise, jury decisions, and the injured party's propensity to sue.

Future research will be focused in two areas. First, using the same or similar data, research should attempt to predict the severity of a loss, given that a loss has occurred. In this study, we used balanced samples of 250 observations with losses and 250 cases without losses. Removing all of the no-loss observations, and concentrating solely on the losses may improve the results. The resulting model may be of interest to both underwriters and claims adjusters.

Second, the long-term possibility of an automated neural network system which could do more than simply accept or reject a policy; but cold also set the appropriate premium, based on the level of risk being accepted by the insurance company. Such an automated system could drastically reduce traditional costs associated with underwriting [28].

References

[1] Gibb, D. E. W. (1972). Lloyd's of London: a study in individualism. London: Lloyd's.
[2] Golding, C. E., & King-Page, D. (1952). Lloyd's. (1 ed.). New York: McGraw-Hill.
[3] Webb, B. L., Harrison, C. M., Markham, J. J., & Underwriters, A. I. f. C. P. C. (1992). Insurance operations. (1st ed.). Malvern, Pa.: American Institute for Chartered Property Casualty Underwriters.
[4] Malecki, D. S., & Underwriters, A. I. f. P. a. L. (1986). Commercial liability risk management and insurance. (2nd ed.). Malvern, Pa.: American Institute for Property and Liability Underwriters.
[5] Hecht, J. (1995). Competition and Technological Change in the Personal Automobile Insurance Industry. CPCU Journal, 48(4), 240.
[6] Tauhert, C. (1998). For Underwriting, NC Blue Turns to an Expert. Insurance and Technology, 23(3), 27.
[7] Trencher, M. L. (1998). Expert Systems Come of Age to Help Drive Insurers' Business. National Underwriter, 102(34), 3.
[8] Daniels, S. (1997). Insurers Seen Lagging in Technology. National Underwriter, 101(34), 5.

[9] Kitchens. (2000). Using Artificial Neural Networks to Predict Losses in Automobile Insurance. Dissertation, The University of Mississippi, Oxford.
[10] Hawley, D. D., Johnson, J. D., & Raina, D. (1990). Artificial Neural Systems: A New Tool for Financial Decision-Making. Financial Analysts Journal, 46(November/December), 63-72.
[11] Wilson, R. L., & Sharda, R. (1994). Bankruptcy Prediction Using Neural Networks. Decision Support Systems, 11, 545-557.
[12] Huang, C. S., Dorsey, R. E., & Boose, M. A. (1995). Life Insurer Financial Distress Prediction: A Neural Network Model. Journal of Insurance Regulation, 3(2), 131-167.
[13] Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. Management Science, 38(7), 926-947.
[14] Barney, D. K., Graves, O. F., & Johnson, J. D. (1999). The Farmers home Administration and Farm Debt Failure Prediction. Journal of Accounting and Public Policy, 18, 99-139.
[15] Lee, T. H., White, H., & Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series model: a comparison of neural network methods and alternative tests. Journal of Econometrics, 56(3), 269-290.
[16] Funahashi, K. (1989). On the Approximate Realization of Continuous mappings by Neural Networks. Neural Networks, 2, 183-192.
[17] Dorsey, R. E., Johnson, J. D., & VanBoening, M. V. (1994). The use of Artificial Neural Networks for Estimation of Decision Surfaces in First Price Sealed Bid Auctions. Netherlands: Kluwer Academic Publishers.
[18] Crane, H. R. (1950). Principles and Problems of Biological Growth. The Scientific Monthly, LXX (6), 376-386.
[19] Nygard, K. E., Ficek, R. K., & Sharda, R. (1992). Genetic Algorithms: Biologically Inspired Search Method Borrows Mechanisms of Inheritance to Find Solutions. OR/MS Today (August), 28--34.
[20] Dorsey, R. E., Johnson, J. D., & Mayer, J. D. (1991). The Genetic Adaptive Neural network Network Training (GANNT) Algorithm for Genetic Feedforward Artificial Neural Systems. Working Paper, The University of Mississippi.
[21] Ramasastry, Amita (2012). "Progressive Car Insurance's 'Snapshot' Experiment: Should Consumers Be Wary of Being Individually Tracked While Driving?" Verdict: Legal Analysis and Commentary from Justia, August 14, 2012.
[22] Klimasauskas, C. (1992). Applying Neural Networks. In R. R. Trippi, and Turban, E. (Ed.), Neural NEtworks in Finance and Investing: Irwin.
[23] Light, R., Singer, J., & Willett, J. (1990). By Design. Cambridge, MA: Harvard University Press.
[24] Stevens, J. (1992). Applied Multivariate Statistics for the Social Sciences. (2 ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
[25] De Fabritiis, C., Ragona, R., & Valenti, G. (2008). "Traffic estimation and prediction based on real time floating car data." Intelligent Transportation Systems. 11th International IEEE Conference on Intelligent Transport Systems, IEEE. 197-203.
[26] McMillan, R.J., Craig, A.D., et al. (1998), "Motor Vehicle Monitoring System for Determining a Cost of Insurance." United States patent Number 5,797,134,
[27] Yvkoff, Liane, (2011). "Gadget Helps Progressive Offer Insurance Discount" Cnet Reviews, March 21, 2011, CBS Interactive Inc.
[28] Davenport, T. H. and J. G. Harris (2005). "Automated Decision Making Comes of Age." MIT Sloan Management Review 46(4): 83-89.